Corpus linguistics with Google?
Stefan Diemer, Saarland University, Germany

Introduction

This paper aims at revisiting the issue whether it is possible to use commercial web search tools such as the Google interface for meaningful corpus research, given recent advances in search technology. It is argued that with the proper methodology these web tools can and should be used for corpus research, since they provide considerable advantages in comparison with both closed corpora and web-based linguistic search tools. Previous investigations of this issue have been mostly negative. Fletcher (in Hundt et al. 2007) ist just one of many citing procedural problems, a criticism that is illustrated by several web-based sample studies (Rosenbach, ibid.). Other researchers advocate the use of specific linguistic search tools like WebCorp, while admitting that these are less efficient than their commercially developed counterparts. More general issues are the question of representativeness and reliability of results (Leech, ibid.). It is not surprising, therefore, that there are few previous studies using a Google-based corpus (among them Lindvall 2004).

1. Key problems with web-based approaches v. closed corpora: corpus size and data organization

- Problem 1: Corpus size: How large is the web, anyway?

It is impossible to say for sure, which means: we cannot give a statistical basis. Some numbers are given, but they are widely fluctuating. However, since the analysis performed here is mostly looking at qualitative and relational data, it may be possible to get reliable results even without knowing the precise size of the corpus.

- Problem 2: Data organization

Mark Davies sums up the problems of web-based corpora by listing searches you can't do with Google, since the latter is not optimized for linguistic use:

- Looking at differences between different styles or types of English (only domains)
- Measuring changes over time
- Grammar-based searches, such as part of speech or lemma
- Wildcard-based searches
- Semantically-based searches
- Finding the word when you don't know what the word is
- Searching for strings of words

(quoted from Davies, Mark 2011 at <http://view.byu.edu/coca/compare-google.asp>)

All these points are essentially correct. But if the research is primarily qualitative and not quantitative, no separate corpus matches the web for providing examples, especially when focusing on non-standard use. Not even the brand-new Google Book Corpus or the Birmingham Blog Corpus is sufficient here, as my examples will show.

So, has the time come to use Google as a corpus tool? I have briefly discussed the shortcomings. But even in its most basic configuration, as Search and Blog Search, the tools provided by Google can be used in corpus linguistics. There are some additional new features that I will discuss at the end of the paper. Possible applications in lexis and phraseology include the quantitative and qualitative documentation of lexical or phraseological innovation in combination with a geographical component and the investigation of word and spelling changes and the use of non-standard features.

2. Google based-lexical study: The return of the prefix? New verb-particle combinations in blogs

I the following example, I will explore how verb-particle combinations, one of the most productive segments of English word-formation, have changed with the advent of online real-time short communication forms such as blogs or their more sophisticated social networking or microblogging varieties like Facebook and Twitter. To illustrate the advantages of a Google-based search, the results are then contrasted with two closed corpora, Mark Davies' Google Book Corpus and the brand-new closed blog corpus by Birmingham University.

Recently I found a nice quotation by an anonymous blogger:

"i ongo, you ongo, he ongoes: it started with ongoing, but is now pandemic. today's annoyance, […]l, downplay. what is wrong with 'play down'? why do we need to keep inventing new verbs to say things we can say perfectly well ...[1]"

2.1. Short history of particle verbs in English and modern tendencies

One of the main trends in the development of English is the long and seemingly unstoppable rise of verb-adverb combinations and the accompanying decline of the prefixes, especially during the Middle English period. As a result, modern English has only very few productive verb prefixes left, in contrast to other languages such as German, where prefixed verbs are much more common and remain productive. This comparatively stable situation may be changing. Many linguists in the 20th century, for example Kennedy (1920) and Kurylowicz (1964) proposed a cyclical development of morphosyntactic verb structure: Phases of free positioning of the particle alternate with phases of complete frozenness (to use Fraser's term), e.g. in the form of prefixes. So far, however, there has been no indication of a cyclical trend. In contrast, the distribution of morphosyntactic verb-particle combination types has been essentially stable over the last 600 years: the prepositional verbs dominant with more than 80%, the adverbial verbs with 15% and the remaining 2,5-5% prefixed verbs. But is this still true under the influence of the WWW, especially since 2000 with the advent of new and pervasive forms of online communication? It is difficult to imagine that English, which is extremely flexible in terms of lexis and semantics, should remain so conservative in the field of verb morphology. After all, prefix verbs are comparatively easy to handle syntactically (just think about the problem of positioning a separable particle) and elegant, even if they are rather unusual in modern English. An early candidate for their use would be the innovative language of information technology. There is some evidence for an increased prefix use in early computer corpora: *downlink*, *uplink* and *throughput*. If one looks at web-based communication, however, there are numerous other examples. Interestingly enough, not only the few remaining productive prefixes (essentially, *down-*, *up-* and *under-*) are used to coin these new terms, but any particles at hand.

Is there any corpus-based evidence of an increase in the use of prefixed verb forms? To answer that question, I examined blogs and microblogs as the most salient web-based communication formats.

2.2. Characteristics of real-time web-based communication forms

Blogging is variously labeled as information and communication technology (ICT) language, computer-mediated communication (CMC), collaborative written or semi-oral discourse. In a recent study on the use of blogs in language teaching, van Compernolle and Abraham note that "formal accuracy is often of little concern to blog authors in many contexts." (Van Compernolle et al. 2009: 209). This is rather understated in view of the plethora of non-standard features blog texts can

[1] Blog *Making it up* - <http://liveotherwise.co.uk/makingitup>

contain in the areas of lexis, syntax and spelling. The more immediate quality of discourse is also variously commented upon. Dylan Glynn (in Evans 2009: 99) discusses the "quasi-spoken language of the blog-diaries", as do most authors in Herring et al. (2007), also noting high innovation and playful use of language. If any innovative use of verbs is to be observed apart from spoken innovation, it should be in this communication medium. This is especially true for the so-called microblogging services, most prominently via Facebook and Twitter.

2.3. Corpus and method

Since the main purpose of the analysis was to provide samples of non-standard use, I decided to use web-based research tools rather than constructing a traditional corpus. No tagging is needed – prefix verbs are comparatively easy to find even in untagged corpora by using simple wildcard searches. This is what Fletcher calls a 'hunting' approach to the 'Web as corpus' or WaC (Fletcher 2007, in Hundt et al.: 28). For the purpose of this investigation, I selected *in* and *on*, the two most frequent particles from the Helsinki and FLOB Corpora (described in detail in Diemer 2008), with a percentage of more than 40% of all particles. Not surprisingly, there are almost no instances of prefixed use in a modern English corpus such as FLOB. These two particles were then manually examined for prefixed use via Google Blogs, Technorati, Twittorati and BlogScope with the 15 most common verbs in the Oxford English Corpus as described by Oxford University Press:[2] *be, have, do, say, get, make, go, know, take, see, come, think, look, want, give.* I also included blogs with non-native user background, since the distinction becomes more and more complex and futile, as David Crystal (2003: 182f.) remarks in his discussion of English as global language. To eliminate hapax occurrences, I will only list the most frequent ones found here (> 1000 uses)

2.4. New prefixed verbs found during the investigation

2.4.1. Verbs with the prefix *in*

In is used in more than 30% of all verb-particle combinations, both diachronically (in the Helsinki Corpus) and synchronically (in the FLOB corpus). In previous analyses (Diemer 2008), there is no prefixed use other than in fossilized forms (e.g. *incoming*) or borrowings from Latin / French (such as *incur*, *invest*) in modern standard English. A blog search, however, produces several innovative prefixed verbs with *in*:

To inbe: This unlikely verb does not exist as finite form, but it is frequent (1391 uses) in its participle (present or past) form, as described in the following example

(1) *19 May 2009 by 1139773367@qq.com: The big shortcoming i have i think is that i can't identify the inbeing of one person from the performance of one's daily life*

To intake: The verb is frequently used transitively, in the sense of 'ingest', 'inhale', 'swallow' (not quite the same as *take in*), also figuratively ('adapt', 'adjust'. As *intake* (verb), it is probably a backformation from the noun *intake*. The verb is fully functional.

(2) *12 Jul 2008 by butifulyletdown: she intakes a sharp breath at his words.*
(3) *11 Apr 2010: He is becoming unhealthy with the stress of work and the food he intakes.*
(4) *14 Feb 2010: This depends on how serious he intakes the acceptance from his peers.*
(5) *8 Aug 2009 by Fiction Theory: So their systems have adapted to strip out the hemoglobin from the intook blood.*

2.4.2. Verbs with the prefix *on*

[2] The Oxford English Corpus at <http://www.askoxford.com/oec/mainpage/oec02/?view=uk>

To ongo: The use of the present participle *ongoing* is, of course, lexicalized and frequent. This might facilitate the formation of other verb forms, and there is evidence that this indeed is happening. Consider the following blog quotation (the source is given above):

(6) *3 Nov 2005 by William Rassman, MD: hi am 27 and went to advanced hairstudio 2 years ago to ongo the laser treatment with the minoxidol treatment*
(7) *10 Nov 2009 by Daniel Of The Boustrophedonical Perspective: The saga still ongoes, but this is another story.*
(8) *8 Dec 2005 by Bill Paley: so, when she evinced a distinct lack of interest in the passavoys and what was going on at their home, it eased an ongoing process, to the point that it onwent much more rapidly.*

To oncome: Often used synonymous with *approach*, but also with *come on*. Fully functional, with more than 400 000 instances of oncoming, almost all negatively connoted. Interestingly, its use can even be analogous to the German 'ankommen' (*to arrive*). The verb is one of the most versatile new coinages found.

(9) *22 Sep 2006 by ON Point: i would be amazed if anyone was alive in the traffic to oncome.*
(10) *21 Nov 2009 by staff: Slacker Radio Mobile application now oncomes to T-Mobile USA BlackBerry Curve 8520.*
(11) *7 Nov 2009 by bclnews.it: Confirmed Habana 6110 under Tirana by // 9600, and at 0101 after all the other RHCs oncame, 6060, 6120, 6140.*
(12) *8 Nov 2005 by lordofcardboard: then oncame carnage and shredded the place to bits.*
(13) *25 Feb 2010 by Geo: Ron also asked: And to further the rights of a shoulder-traveling cyclist, a vehicle in the oncoming lane of a two-lane road should not overtake another vehicle in the oncoming lane if there is bicycle traffic in the opposite direction*
(14) *21 Feb 2010 by JJS: Frolicsome pooches sense oncoming spring.*
(15) *21 Mar 2009 by vkpgqkvexqcznw: Oncoming to you live from Hong Kong!*

To onlook: The verb is very frequent, with more than 9000 uses for *onlooking* alone. There are some punning uses of *to onlook*, analogous to the existing *onlooker*. But the prefix verb is also used in the sense of 'watch', 'witness' and thus more flexible than *look on* in this context.

(16) *29 Sep 2009 by theory friction practice: The performer's blindness made it easier for onlookers to onlook.*
(17) *26 Nov 2009 by John Cole: More cop cars have arrived, and some neighbors are beginning to onlook, they tell me*
(18) *24 Feb 2010 by admin: It has a charming thatched Bar with Pool table and DSTV with food and drink refreshments on sale - the bar onlooks a gorgeous circular swimming pool which is open to all –.*
(19) *11 Jan 2010 by k_ANNE: It's so hard to have faith sometimes in a world that onlooks with doubt.*
(20) *3 Oct 2009 by relicpro: Wedding Makeup. Portland, Oregon - The flower girl onlooks as the bride puts on her makeup for her wedding at Old Laurelhurst Church in SE Portland.*
(21) *24 Jan 2010 by Fruitarian Mango: The Malanda home has a great rear view of the valley, but no established fruit trees that I noticed, and it's onlooked by distant surrounding properties.*
(22) *26 Jan 2008 by Nirnimesh: as a child, i actually used to enjoy this day primarily because it was a holiday but also because of the colorful jhanki (processions) that different stats would march near the red fort, and the president onlooked them.*
(23) *22 Jan 2008 by POD: wigs hurt not just the onlooker but the onlooked.*

2.4.3. Some other prefixed verbs found during analysis

To aset: This verb form (Google Blogs: 253 occurrences) is an exciting revival of the OE verb *asettan*: 'set, set up, to set out on (journeys)'. The meaning found in Google Blogs is closest to *set up*. However, examination of the corpus results shows that many occurrences have a different origin, namely spelling errors: *Aset = a set*; a spelling error due to missing line break or *Aset = set*; a spelling error because the keys a and s are next to each other on ASCII keyboards.

(24)*11 Feb 2010 by sfag30: doh - any advice on how to aset a vpn then.*

To atstand: The verb is very rare, but used analogous to the Old English *aetstandan*: 'stand fixed':

(25)*1 Aug 2009 by oops: that they come in, all guests atstand up together, jue-ming master busy smiling salute to the people together even after the people led him straight back to them that the new platform.*

To forespeak: The verb is quite common in blogs:

(26)*4 Sep 2009 by John: so allow doom to use his unquestioned powers of prediction to forespeak the coming year of competitive feats.*

To offput: The verb is not in the OED, but Google Blogs has more than 1900 occurrences in blogs since 2000, which makes it very common. Still, that is not much compared with the more than 119 million times its adverbial equivalent *put off* is used in the same time span. Syntactically, *offput* works like a full verb: there is active and passive, transitive and intransitive use and negation:

(27)*12 Jul 2008 by Michael: The only thing I'm still a little offput by is Jacobs' statement that EA had nothing to do with this decision.*

2.5. Results of the investigation

The analysis shows that there is considerable innovation in prefix verbs used in blogs. This is, in itself, remarkable, because the innovations are almost all considered non-standard. The widespread use of non-standard forms in blogs indicates that this form of discourse is considerably more flexible than other written discourse types. However, in most cases, the proportion of prefix verbs is minuscule compared to the more conventional adverbial or prepositional verbs. This points to a very early stage of innovation. There are some notable exceptions, like *forespeak*, *oncome* and *offput*. Several possible reasons for the use of non-standard verb forms can be proposed:

- Non-native language use
- Analogy with existing prefix verbs
- Special-purpose use
- Playful use of language
- Facilitation of syntax

Reason 1: Non-native language use: A strong contributing factor may be the use of English as second language, which means that patterns of speech from the original language can be transposed onto English, as in the example *inknow*:

(28)*1 Oct 2008 by kingu: grandpas come from canton, but, i inknow above lund styles, schools or sect. please help me, yep!*

This verb use seems to originate in a language that uses prepositioned particles to negate the verb, such as Chinese *bù*. It could be argued that this flexible use of negative prefixes will become more widespread as a consequence of increased use of globalized English.

Reason 2: Analogy with existing prefix verbs: A topic of ongoing discussion is the question why some prefixes (like the one in *ongoing*) are still productive with some verbs, while most are not. There is also no clear reason why many of the standard prefix verb forms are incomplete, such as *incoming* (but not *to income*) or *outgoing* (but not *to outgo*). I cannot give a reason here, but the analysis shows that these left-over prefix forms are not only in frequent use, but that they motivate the formation of analogous forms. *Oncoming* is created according to the pattern established by *incoming*:

(29)*9 Mar 2010: […] pulling a friend from oncoming traffic on a busy road.*

Reason 3: Special-purpose use: Another reason for the use of non-standard prefix verbs clearly is the specialized background of the blog writer. The special-purpose environment facilitates the formation of terms that could not be used in a general context and carry a precise, complex and limited meaning. Inbeing (for 'the essence of one's self') is used as a complex psychological term, while *inhave* and *intake* are used in a medical context. *To inlet* is introduced as a wood- or metalworking term, *to infall* is astronomy jargon.

Reason 4: Playful use of language: This is, of course, the classical reason for making up new words. Connie Eble (1996) comments on this feature in her insightful analysis of college talk. She notes that this mostly happens in per group communication. A good example is *indone*, which a female blogger used instead of *done in*, addressing a female peer group. Others are *incame* and *inthought*. Many new prefix verbs also start off as proper names for new companies or products, such as *onbeing*, *onthink* and *onthought*. Here, the innovation is driven by marketing considerations: a new, fresh name for a new, innovative product.

Reason 5: Facilitation of syntax: Remarkably, many prefix verbs seem to be used because they simplify the syntax. Most commonly, adverbs are replaced by prefixes, undoing the shift away from prefix verbs in Old and Early Middle English. For example *inkick* and *inbring* are used instead of *kick in* and *bring in*, avoiding the question of where to position the adverbial particle. A similar motivation can be seen in intake, as in 'she intakes a sharp breath at his words'.
The search for easier syntax can lead to interesting alternatives, such as replacing a negation by a negative prefix. Thus, *I inknow* is used instead of *I didn't know*. The facilitation of syntax may be accompanied by a shift in cognitive perception or, to put it structurally, a re-strengthening of the cohesion of particle and verb. Consider *ontake* and *ongo*, two non-standard prefix verbs used instead of take on and go on. The phrase *in order to ontake such an insane project* is, syntactically, more compact than *in order to take on such an insane project* or even *in order to take such an insane project on*. It has the added advantage of rejoining the two components of the verb phrase. The one-word phrase in the first example is easier to process from a cognitive perspective, since the prepositional access point (the main determiner of meaning according to the cognitive approach) precedes the verb. As primary vehicles of embodied meaning, particles may be perceived as easier to understand. With the plethora of post-positioned particles in existence, accessing the meaning of the particular phrase has to be done via the verb, then the particle. This may be more difficult, leading to the search for an alternative. The factor time should be mentioned, since I would argue that both bloggers and readers of blogs aim at maximizing output and intake, respectively.

2.6. Further research

This study is just a snapshot of a couple of new non-standard prefixed verbs in order to demonstrate that there is considerable potential for research in this field. I wanted to show a very early stage of

this innovation, something that only the use of the web as a corpus could provide. The obvious shortcoming of this study is the lack of a clear quantitative element, for reasons discussed above. An investigation of additional prefixed forms on the basis of a larger questionnaire and a more precise quantification seem the logical next steps. Hopefully this can be done quite soon. For the first time in almost 1000 years, verb prefixation may be increasing again. This has the potential to fundamentally alter the character of the English language and to reverse a long-time trend.

2.7. Comparison with closed-corpus searches

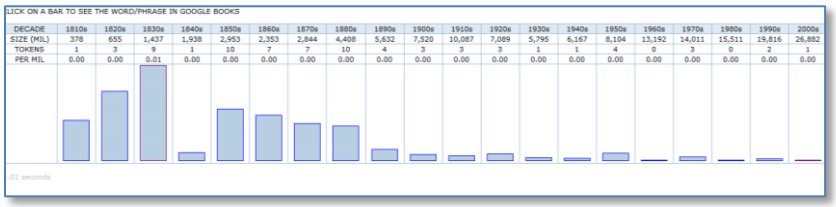
2.7.1. Google Book Corpus

Let's now contrast these findings with what we could have gained from a closed corpus. Just recently, Mark Davies published his Google Books (American English) Corpus, with 155 billion words, a staggering number which makes it one of the biggest closed corpora ever next to Googles n-gram corpus itself. The result of an investigation after the same pattern as described above provides the following results:

There are (after laboriously excluding name and other mismatches) no matching strings for

to inbe,
to intake,
to ongo,
to oncome,
to onlook,
to aset
to atstand,
to forespeak
to offput

There is one interesting match, however.



(Search for *inbeing*, source: Google Book Corpus at BYU)

There are a few scattered instances of *inbeing*, well below the minimum percentage threshold. This seems to indicate that there is, indeed, a rarely used pattern (limited to philosophy) that can now be used to a wider extent in online communication again.

2.7.2. Birmingham Blog Corpus (BBC)

You could argue that the negative results stem from the fact that the Google Book Corpus does not contain the discourse type in question, blogs, at all. I am therefore particularly grateful to Andrew Kehoe, Matt Gee and their colleagues at Birmingham City University for just recently making the 100-million-word Birmingham Blog Corpus available for use and publicly searchable through WebCorp (WebCorp interface at <http://wse1.webcorp.org.uk>). There are, indeed, some limited results for the forms I looked at:

No results for:
to inbe
to ongo
to atstand
to forespeak

to intake,
4 instances (only as infinitive)

1:	content they choose	to intake	? There can be no
2:	it take multiple reads	to intake	the sense of it
5:	fans with no air	to intake	, that's a clue. That's
7:	we just don't get	to intake	information firsthand.

The results are similar to the ones found in the open web search, but vastly less numerous and much less varied. In addition, the tagging process tagged all verbs as nouns, thus negating the theoretical advantage of a POS search.

to oncome
533 instances of *oncoming*, but no other forms

to onlook
10 instances of *onlooking*, no other forms

to aset
1 result, tagged as a noun

1:	tables and a podium	aset	up on the small

to offput
221 instances, all participle:

The results from the BBC show that the discourse of blogs seems to be the area where this innovation occurs. In addition, the limited results (only infinitive and participle forms, respectively) indicate that even a specialist corpus will only produce examples after they have been conventionalized to a large extent.

2.7.3. Results

The comparison of results from open and closed corpora seem to support the theory that you cannot rely on closed corpora to really find language innovation as it happens. In order to discuss trends, it is, in my opinion, necessary , to use the Web as Corpus / hunting approach for this dynamic and innovative discourse type.

3. Other applications

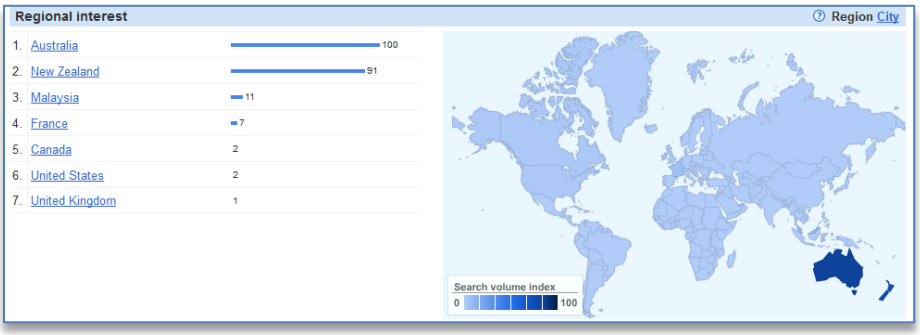
3.1. New research tools from Google

Basic Web- or Google-based searches also have potential uses in other areas of linguistics, where the focus is on lexical and qualitative features. But there are also some newer features that may make using the web as a corpus increasingly possible and attractive. The previous mostly negative verdicts may have to be reconsidered in view of powerful tools recently developed by Google, specifically “Trends” and “Insights for Search”. Both use Google’s infrastructure for web search, but provide considerably more data than the standard interface. “Google Trends” gives researchers the option of

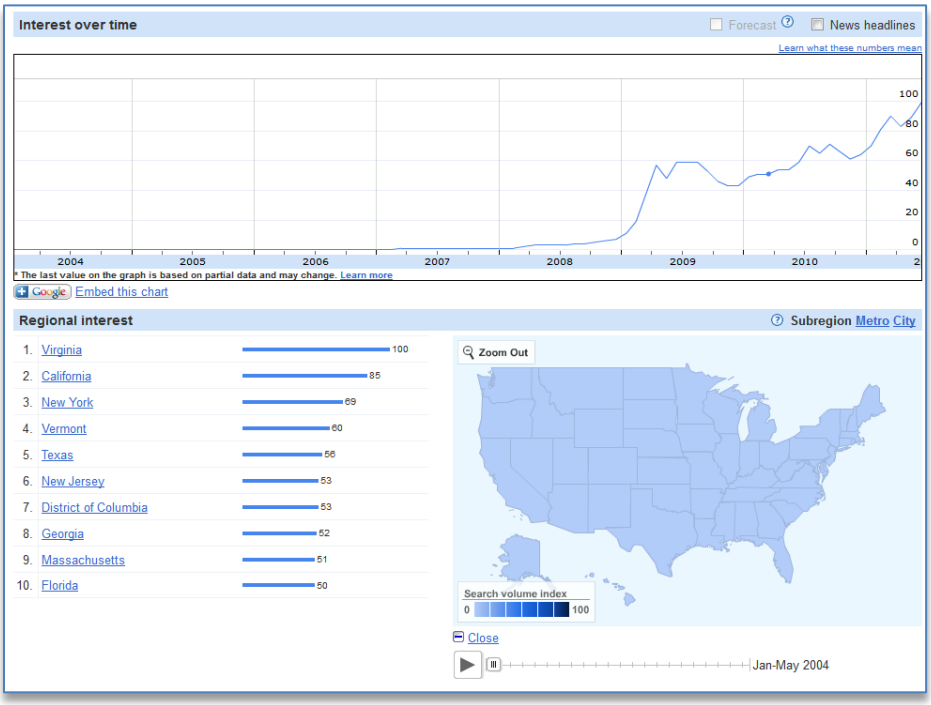
instantly examining search volumes for specific search terms or phrases, either diachronically or synchronically. The data is provided in numerical, graphical and normalized formats and can be adjusted for region, country, city or language. It is possible to plot developments in lexical or phrase search use and to compare search terms, using different variables such as geographical or diachronic distribution. “Insights for Search” uses the same data as “Trends”, but is aimed at researchers (or advertisers) and includes several additional features, including geographic “heat maps” that illustrate the dissemination of search terms, and a forecast option that calculates the potential future use of search terms. The inclusion of blogs and other online text types in the corpora allows the examination not only of past, but also of ongoing language innovations and changes.

3.2. Sociolinguistics

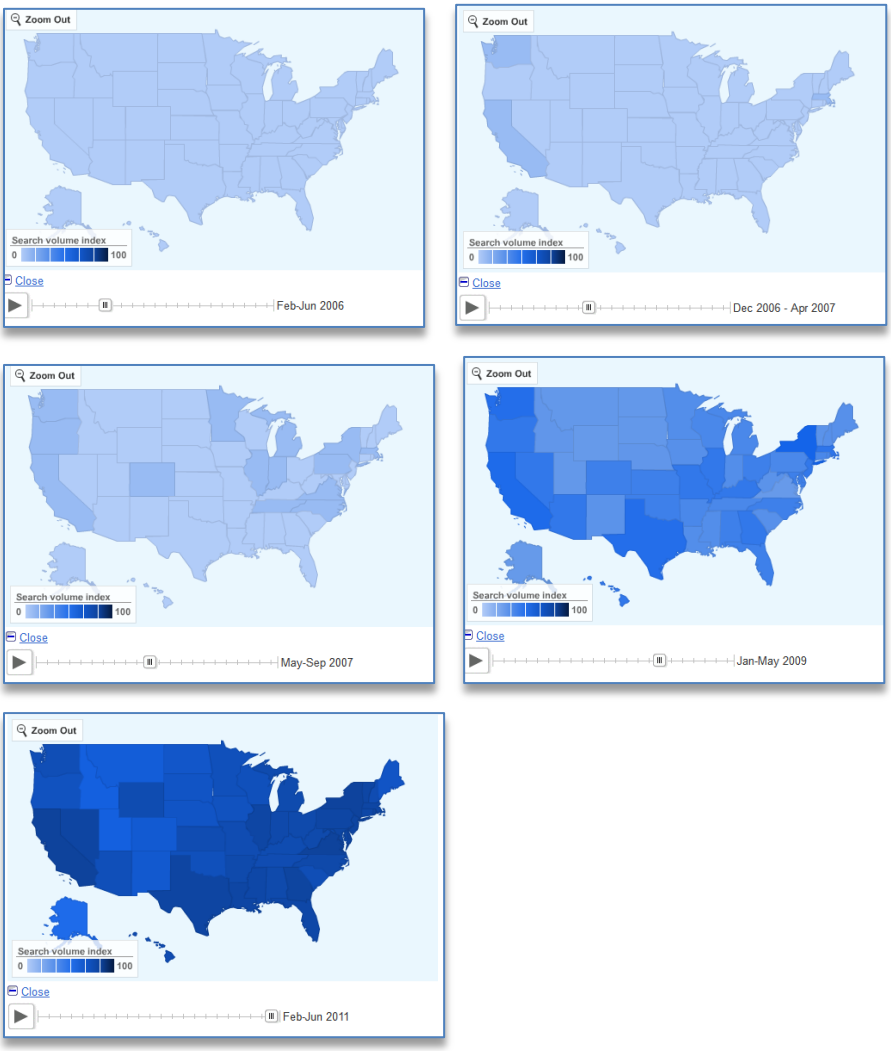
In sociolinguistic research, the tools allow the precise description of sociolects based on real-time data. Again, the quantitative element here is less important than the actual documentation of the use and spread of innovation. Keeping in mind the limitations of the Web as Corpus approach discussed above, one example, again, is the spread of new lexical variants. In addition, “Insights provides exciting (if unspecified) geographical data:



Google Insights search for *kindy*



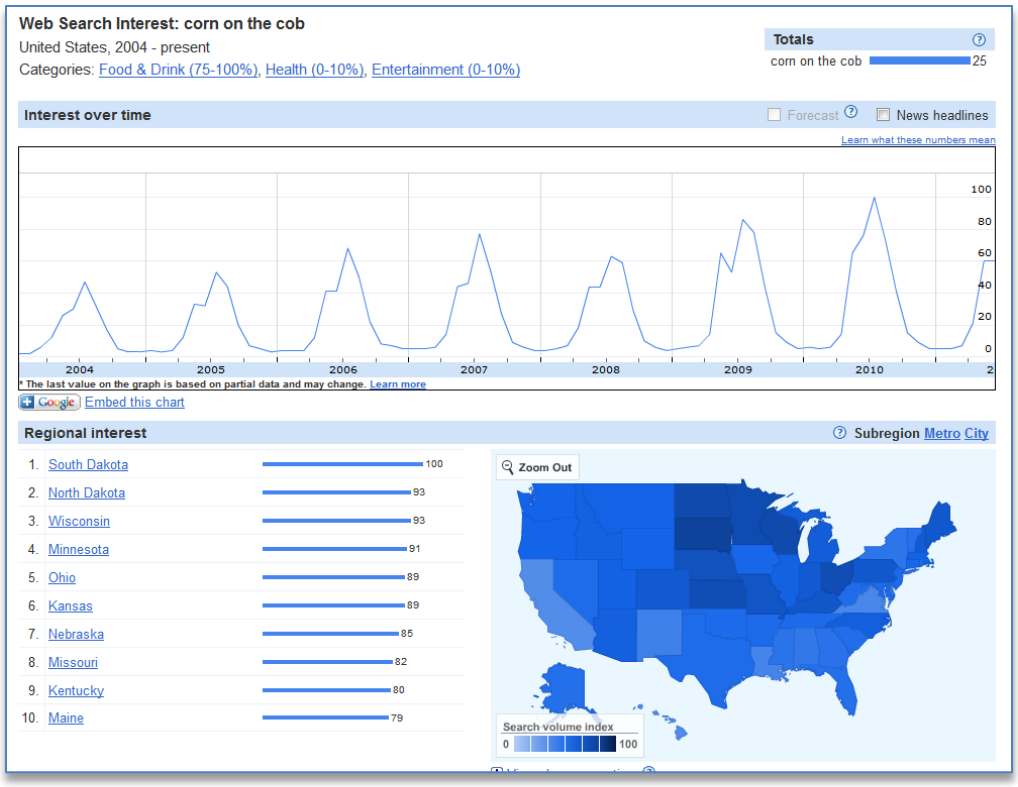
Google Insights search for *twitter* (Interest over time graph)



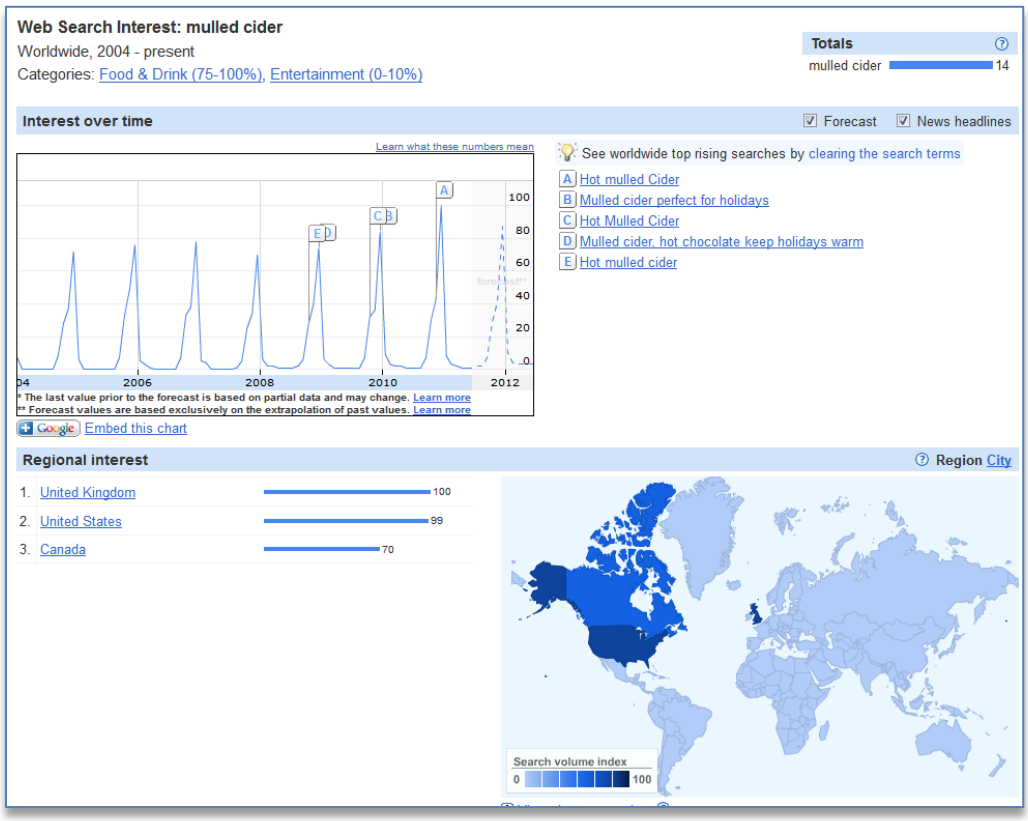
Google Insights search for *twitter* (Dynamic interest over time)

3.3. Cognitive Semantics

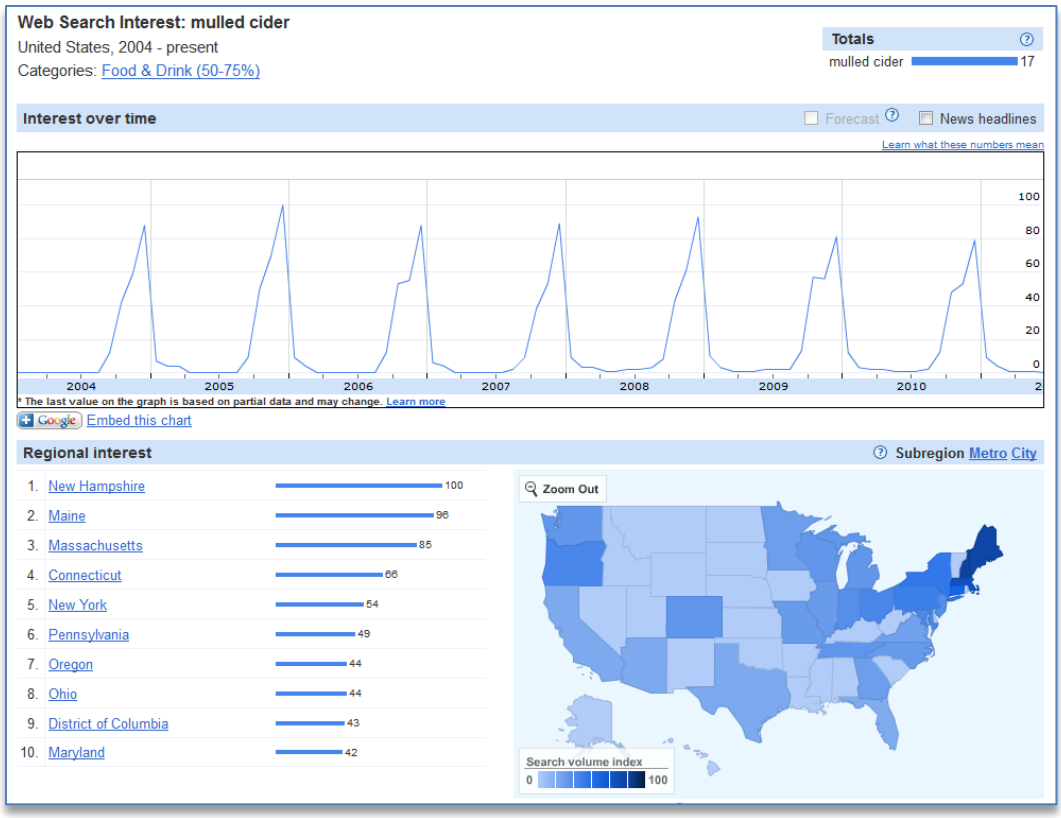
In cognitive semantics this approach can pinpoint connotations much more convincingly than existing corpus-based methods. Again, the focus here lies, due to the restraints discussed above, in close syntactic patterns like phrasal verbs. But also framing and cultural backgrounds can be analyzed, as the following examples show:



Google Insights search for *corn-on-the cob* (US)



Google Insights search for *mulled cider* (World)



Google Insights search for *mulled cider* (US)

4. Conclusion: Should we use Google for Corpus Linguistics?

As these few examples have shown, these web tools can and should be used for corpus research, since they provide considerable advantages in comparison with both closed corpora and web-based linguistic search tools. There is one grave disadvantage: you can‘t give clear numbers. In corpus linguistics, this is traditionally seen as very bad. But let me put forward a slightly heretical idea here: do we really need all these numbers? Remember James Fillmore’s joke about corpus linguists and armchair linguists? Or Noam Chomsky’s dismissal of the number-crunchers? Well, he may have had a point.

In a seminal talk at the recent ICAME 32 conference in Oslo, Michael Stubbs from Trier University pointed out key problems in the mostly number-based approach that CL has been pursuing. It got me thinking. In the process of having become a fully-fledged discipline , it is, I think, time for us as corpus linguists to accept that it is not all about numbers anymore, that we should move towards a more “grown-up” approach, finally integrating empiricist theory and rationalist elements. For too long we have ignored the centuries-old dichotomy between these two approaches. In the last 20 years and especially today we have become a bit mesmerized by the rapid advances in corpus technology. Looking back to the minuscule Brown corpus in the 1960s it is easy to say: “bigger,faster, better”. But it is not enough to just collate our concordances and put together our closed corpora. David Crystal said two weeks ago that our view of diachronicity and synchronicity may be breaking down with the online environment of blogs, where you can change and mix posts even years after first publishing them. I think the age of closed corpora is coming to an end rather sooner than we think. So why not use the fastest existing algorithm to search the full field of online discourse, extracting some quantitative, but mainly qualitative features and then discussing them from a qualitative rather than a primarily quantitative perspective.

After all, Francis Bacon, the prototypical empiricist, asked us to search for "fingerposts" (instantiis cruciis) in search of new knowledge, rather than listing (and adding) empiric results. Kant, in trying to integrate the two directions, finds harsh words for a purely empiric as well as a purely rationalist approach. He counsels that "Rational thought needs to encounter nature in order to learn, but not like a student, purely receptive, but like a judge, forcing a witness to answer its questions." (Kant 1787: 4). My former university teacher and mentor Peter Erdmann, recently retired at Berlin Technical University, had a phrase that he would ask all corpus linguistics students when they presented their flawless statistics to him. He would look them in the eye and ask: "But what does it mean?". I think we have to try and answer that question. Numbers are important, but linguistics is not a natural science, it is and will remain an art. If we just focus on closed corpora and numbers, we will be missing the bus on innovation.

For this purpose, and to that extent, yes, Google is a suitable corpus tool.

References

- Abraham, Lee B. & Lawrence Williams (eds.). 2009. *Electronic discourse in language learning and language teaching.* Amsterdam: John Benjamins.
- Bacon, Francis. 1660. Novum organum scientiarum. 2nd ed. Amstelodanum: Johann Ravestein.
- Bansal, Nilesh & Nick Koudas. 2007. Searching the Blogosphere. In *Proceedings of the 10th international Workshop on Web and Databases*. Beijing: WebDB 2007.
- BlogScope at <http://www.blogscope.net> (31 May 2011)
- Van Compernolle, Rémi A. & Lee B. Abraham. 2009. Interactional and discursive features of English-language weblogs for language learning and teaching. In: Lee B. Abraham, Lawrence Williams (eds.), *Electronic discourse in language learning and language teaching.* Amsterdam: John Benjamins: 193-212.
- Crystal, David. 2008. *Txtng: the Gr8 Db8*. Oxford: Oxford University Press.
- Davies, Mark. 2011. "The Corpus of Contemporary American English (COCA) and Google / Web as Corpus." Full text at <http://view.byu.edu/coca/compare-google.asp> (31 May 2011)
- Diemer, Stefan. 2008. "Das Internet als Korpus? Aktuelle Fragen und Methoden der Korpuslinguistik." (The internet as a corpus? Current questions and issues in corpus linguistics"). Saarland Working Papers in Linguistics 2 (2008): 29-57. Full text at <http://scidok.sulb.uni-saarland.de/volltexte/2009/2148/> (31 May 2011)
- Diemer, Stefan. 2008. *Die Entwicklung des englischen Verbverbandes – eine korpusbasierte Untersuchung.* ("The development of verb-particle combinations in English – a corpus-based study.") Habilitationsschrift. Berlin: TU Berlin.
- Diemer, Stefan. 2010. "It's all a bit upmessing. Non-standard verb-particle combinations in blogs." In: Saarland Working Papers in Linguistics 3 (2009): 35-56. Full text at < http://scidok.sulb.uni-saarland.de/volltexte/2010/3417/> (31 May 2011)
- Eble, Connie. 1996. *Slang and Sociability*. Chapel Hill: University of North Carolina Press.
- Evans, Vyvyan & Stéphanie Pourcel (eds.). 2009. *New directions in cognitive linguistics.* Amsterdam: John Benjamins.
- Fletcher, William H. 2007. Concordancing the web: promise and problems, tools and techniques. In: Corpus Linguistics and the Web. Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds), 25-46. Amsterdam: Rodopi.
- Giltrow, Janet & Dieter Stein (eds.). 2009. *Genres in the internet*. Amsterdam: John Benjamins.
- Glynn, Dylan. 2009. Polysemy, syntax and variation: A usage-based method for Cognitive Semantics. In: Vyvyan Evans, Stéphanie Pourcel (eds.), *New directions in cognitive linguistics.* Amsterdam: John Benjamins: 77-106.

- Google Blog Search at <http://blogsearch.google.com> (31 May 2011)
- Google Books: American English Corpus at <http://googlebooks.byu.edu/> (31 May 2011)
- Google Insights for Search Help centre at <http://www.google.com/support/insights/?hl=en-GB> (31 May 2011)
- Google Trends: About Google Trends at <http://www.google.com/intl/en/trends/about.html> (31 May 2011)
- Herring, Susan C. & Brenda Danet (eds.). 2007. *The multilingual internet: Language, culture, and communication online.* New York: Oxford University Press.
- Hiltunen, Risto. 1983. The decline of the prefixes and the beginnings of the English phrasal verb. Turku: Turun Yliopisto.
- Hundt, Marianne, Nesselhauf, Nadja & Biewer, Carolin (eds). 2007. Corpus Linguistics and the Web. Amsterdam: Rodopi.
- Kant, Immanuel. 1787. Kritik der reinen Vernunft. 2nd ed. Riga: Hartknoch.
- Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In: Corpus Linguistics and the Web. Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds), 133-150. Amsterdam: Rodopi.
- Lindvall, Lars. 2004. Using Google for corpus linguistics : remarks on the usage of the Swedish temporal conjunction "efter det att" 'after' and its equivalents in French, Italian and Spanish. Translation and corpora (2004): 189-207.
- Rosenbach, Anette. 2007. Exploring constructions on the web: a case study. In: Corpus Linguistics and the Web. Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds), 167-190. Amsterdam: Rodopi.
- Twitter at <http://twitter.com> (31 May 2011)
- Twittorati at <http://twittorati.com/> (21 May 2011)

Contact Information

Stefan Diemer
FR 4.3 Anglistik, Amerikanistik
und Anglophone Kulturen
Universität des Saarlandes
Postfach 15 11 50
D-66041 Saarbrücken
Germany